Biomedical Optics EXPRESS

Automated analysis of multimodal fluorescence lifetime imaging and optical coherence tomography data for the diagnosis of oral cancer in the hamster cheek pouch model

Paritosh Pande,[1,*] Sebina Shrestha,[1] Jesung Park,[1] Irma Gimenez-Conti,[2] Jimi Brandon,[2] Brian E. Applegate,[1] and Javier A. Jo[1]

[1]*Biomedical Engineering Department, Texas A&M University, College Station, Texas 77843, USA*

[2]*Department of Carcinogenesis, The University of Texas M.D. Anderson Cancer Center, Smithville, Texas 78957, USA*

[*]*pande@illinois.edu*

Abstract: It is known that the progression of oral cancer is accompanied by changes in both tissue biochemistry and morphology. A multimodal imaging approach combining functional and structural imaging modalities could therefore provide a more comprehensive prognosis of oral cancer. This idea forms the central theme of the current study, wherein this premise is examined in the context of a multimodal imaging system that combines fluorescence lifetime imaging (FLIM) and optical coherence tomography (OCT). Towards this end, in the first part of the present study, the diagnostic advantage obtained by using both fluorescence intensity and lifetime information is assessed. In the second part of the study, the diagnostic potential of FLIM-derived biochemical features is compared with that of OCT-derived morphological features. For an objective assessment, several quantitative biochemical and morphological features from FLIM and OCT data, respectively, were obtained using signal and image processing techniques. These features were subsequently used in a statistical classification framework to quantify the diagnostic potential of different features. The classification accuracy for combined FLIM and OCT features was estimated to be 87.4%, which was statistically higher than accuracy based on only FLIM (83.2%) or OCT (81.0%) features. Moreover, the complimentary information provided by FLIM and OCT features, resulted in highest sensitivity and specificity for the combined FLIM and OCT features for discriminating benign (88.2% sens., 92.0% spec.), pre-cancerous (81.5% sens., 96.0% spec.), and cancerous (90.1% sens., 92.0% spec.) classes.

OCIS codes: (300.2530) Fluorescence, laser-induced; (170.6920) Time-resolved imaging; (170.4500) Optical coherence tomography(170.6935); (100.0100) Image processing; (170.6935) Tissue characterization; (170.1610) Clinical applications.

#261106 Received 15 Mar 2016; revised 18 Apr 2016; accepted 19 Apr 2016; published 25 Apr 2016

References and links

1. N. Howlader, A. Noone, M. Krapcho, N. Neyman, R. Aminou, S. Altekruse, C. Kosary, J. Ruhl, Z. Tatalovich, H. Cho, "Seer cancer statistics review, 1975–2009 (vintage 2009 populations)," Bethesda, MD: National Cancer Institute (2012).
2. L. M. Abbey, G. E. Kaugars, J. C. Gunsolley, J. C. Burns, D. G. Page, J. A. Svirsky, E. Eisenberg, D. J. Krutchkoff, and M. Cushing, "Intraexaminer and interexaminer reliability in the diagnosis of oral epithelial dysplasia," Oral Surg., Oral Med., Oral Pathol. Endodontol. **80**, 188–191 (1995).
3. V. K. Ramanujan, J.-H. Zhang, E. Biener, and B. Herman, "Multiphoton fluorescence lifetime contrast in deep tissue imaging: prospects in redox imaging and disease diagnosis," J. Biomed. Opt. **10**, 051407 (2005).
4. M. C. Skala, J. M. Squirrell, K. M. Vrotsos, J. C. Eickhoff, A. Gendron-Fitzpatrick, K. W. Eliceiri, and N. Ramanujam, "Multiphoton microscopy of endogenous fluorescence differentiates normal, precancerous, and cancerous squamous epithelial tissues," Cancer Res. **65**, 1180–1186 (2005).
5. M. C. Skala, K. M. Riching, D. K. Bird, A. Gendron-Fitzpatrick, J. Eickhoff, K. W. Eliceiri, P. J. Keely, and N. Ramanujam, "In vivo multiphoton fluorescence lifetime imaging of protein-bound and free nicotinamide adenine dinucleotide in normal and precancerous epithelia," J. Biomed. Opt. **12**, 024014 (2007).
6. P. Wilder-Smith, K. Osann, N. Hanna, N. E. Abbadi, M. Brenner, D. Messadi, and T. Krasieva, "In vivo multiphoton fluorescence imaging: a novel approach to oral malignancy," Lasers Surg. Med. **35**, 96–103 (2004).
7. E. S. Matheny, R. Mina-Araghi, M. Brenner, N. M. Hanna, W. Jung, Z. Chen, and P. Wilder-Smith, "Optical coherence tomography of malignancy in hamster cheek pouches," J. Biomed. Opt. **9**, 978–981 (2004).
8. P. Wilder-Smith, T. Krasieva, W.-G. Jung, J. Zhang, Z. Chen, K. Osann, and B. Tromberg, "Noninvasive imaging of oral premalignancy and malignancy," J. Biomed. Opt. **10**, 051601 (2005).
9. P. Wilder-Smith, K. Lee, S. Guo, J. Zhang, K. Osann, Z. Chen, and D. Messadi, "In vivo diagnosis of oral dysplasia and malignancy using optical coherence tomography: preliminary studies in 50 patients," Lasers Surg. Med. **41**, 353–357 (2009).
10. R. Richards-Kortum and E. Sevick-Muraca, "Quantitative optical spectroscopy for tissue diagnosis," Annu. Rev. Phys. Chem. **47**, 555–606 (1996).
11. M. G. Müller, T. A. Valdez, I. Georgakoudi, V. Backman, C. Fuentes, S. Kabani, N. Laver, Z. Wang, C. W. Boone, R. R. Dasari, "Spectroscopic detection and evaluation of morphologic and biochemical changes in early human oral carcinoma," Cancer **97**, 1681–1692 (2003).
12. P. Wilder-Smith, T. Krasieva, W.-G. Jung, J. Zhang, Z. Chen, K. Osann, and B. Tromberg, "Noninvasive imaging of oral premalignancy and malignancy." J. Biomed. Opt. **10**, 051601 (2005).
13. P. Pande, S. Shrestha, J. Park, M. J. Serafino, I. Gimenez-Conti, J. Brandon, Y.-S. Cheng, B. E. Applegate, and J. A. Jo,"Automated classification of optical coherence tomography images for the diagnosis of oral malignancy in the hamster cheek pouch," J. Biomed. Opt. **19**, 086022 (2014).
14. J. Park, J. A. Jo, S. Shrestha, P. Pande, Q. Wan, and B. E. Applegate, "A dual-modality optical coherence tomography and fluorescence lifetime imaging microscopy system for simultaneous morphological and biochemical tissue characterization," Biomed. Opt. Express **1**, 186–200 (2010).
15. X. Tang,"Texture information in run-length matrices," IEEE Trans. Image Process. **7**, 1602–1609 (1998).
16. A. Criminisi, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," Foundations and Trends in Computer Graphics and Vision **7**, 81–227 (2011).
17. J. Demšar, "Statistical comparisons of classifiers over multiple data sets," J. Mach. Learn. Res. **7**, 1–30 (2006).
18. D. G. Farwell, J. D. Meier, J. Park, Y. Sun, H. Coffman, B. Poirier, J. Phipps, S. Tinling, D. J. Enepekides, and L. Marcu, "Time-resolved fluorescence spectroscopy as a diagnostic technique of oral carcinoma: validation in the hamster buccal pouch model," Arch. Otolaryngol. Head Neck Surg. **136**, 126–133 (2010).
19. M. W. Lingen, J. R. Kalmar, T. Karrison, and P. M. Speight, "Critical evaluation of diagnostic aids for the detection of oral cancer," Oral Oncol. **44**, 10–22 (2008).
20. R. Mehrotra and D. K. Gupta, "Exciting new advances in oral cancer diagnosis: avenues to early detection," Head Neck Oncol. **3**, 1–9 (2011).
21. A. Gillenwater, R. Jacob, R. Ganeshappa, B. Kemp, A. K. El-Naggar, J. L. Palmer, G. Clayman, M. F. Mitchell, and R. Richards-Kortum, "Noninvasive diagnosis of oral neoplasia based on fluorescence spectroscopy and native tissue autofluorescence," Arch. Otolaryngol. Head Neck Surg. **124**, 1251–1258 (1998).
22. D. De Veld, M. Witjes, H. Sterenborg, and J. Roodenburg, "The status of in vivo autofluorescence spectroscopy and imaging for oral oncology," Oral Oncol. **41**, 117–131 (2005).
23. I. Pavlova, M. Williams, A. El-Naggar, R. Richards-Kortum, and A. Gillenwater, "Understanding the biological basis of autofluorescence imaging for oral cancer detection: high-resolution fluorescence microscopy in viable tissue," Clin. Cancer Res. **14**, 2396–2404 (2008).
24. R. Drezek, C. Brookner, I. Pavlova, I. Boiko, A. Malpica, R. Lotan, M. Follen, and R. Richards-Kortum, "Autofluorescence microscopy of fresh cervical-tissue sections reveals alterations in tissue biochemistry with dysplasia," Photochem. Photobiol. **73**, 636–641 (2001).
25. J. E. Bouquot and R. J. Gorlin, "Leukoplakia, lichen planus, and other oral keratoses in 23,616 white americans over the age of 35 years," Oral Surg., Oral Med., Oral Pathol. **61**, 373–381 (1986).

#261106 Received 15 Mar 2016; revised 18 Apr 2016; accepted 19 Apr 2016; published 25 Apr 2016

26. B. W. Neville and T. A. Day, "Oral cancer and precancerous lesions," Ca-Cancer J. Clin. **52**, 195–215 (2002).
27. I. Pavlova, K. Sokolov, R. Drezek, A. Malpica, M. Follen, and R. Richards-Kortum, "Microanatomical and biochemical origins of normal and precancerous cervical autofluorescence using laser-scanning fluorescence confocal microscopy," Photochem. Photobiol. **77**, 550–555 (2003).
28. Y. Wu, P. Xi, J. Qu, T.-H. Cheung, and M.-Y. Yu, "Depth-resolved fluorescence spectroscopy reveals layered structure of tissue," Opt. Express **12**, 3218–3223 (2004).
29. Y. Wu and J. Y. Qu,"Combined depth-and time-resolved autofluorescence spectroscopy of epithelial tissue," Opt. Lett. **31**, 1833–1835 (2006).

1. Introduction

Oral cancers have an overall five-year survival rate of approximately 60%. However, when detected in early stages, oral cancers can have a five-year survival rate as high as 80-90% [1]. Early detection of oral cancer, therefore, holds the key for improving survival rates. Diagnosis of oral cancer in late stages can be mainly attributed to the current method of diagnosis of oral malignancies, which is primarily based on a visual assessment of oral lesions, followed by biopsies in suspicious cases. The distinction between a benign and a malignant lesion is, however, not always very clear by a mere visual examination. In the absence of a reliable screening tool, therefore, it is not easy to determine which *abnormal looking* lesions in the oral cavity are worthy of concern. Since it is impractical to perform biopsies on every suspicious lesion, the standard protocol is to watch a lesion over an extended period of time to determine if it is indeed malignant. This often leads to a situation where a potentially malignant lesion, which might be easily treatable at an earlier stage, advances into a later untreatable condition. Even in cases where suspicious lesions are well identified, due to the complex nature of oral lesions, the choice of the optimal site for biopsy remains an important concern. Additionally, high rates of inter and intra-observer variability in histopathological evaluation of oral biopsies further confound the diagnosis of oral cancer [2]. It is known that the progression of a malignant lesion from an early stage to later stages is accompanied by both biochemical changes like alterations in the relative abundances of tissue autofluorophores [3–5] and morphological changes like epithelial thickening and loss of the layered tissue structure [6–9]. An imaging system, capable of providing information about both tissue biochemistry and morphology, could therefore serve as an effective screening tool for possible malignancy of large number of oral lesions, which are routinely encountered by dentists and other physicians during visual examination of the oral cavity. Motivated by this idea, in this study, we primarily seek to examine the premise that using both biochemical and morphological information, obtained from fluorescence lifetime imaging (FLIM) and optical coherence tomography (OCT), respectively, increases the diagnostic accuracy for oral cancer, compared to using only one type of information.

Use of fluorescence imaging for the diagnosis of oral cancer has been reported in several previously published studies. It is suggested that the alterations in the intrinsic fluorescence properties of the oral tissue, which are indicative of the progression of oral cancer, are a consequence of changes in the relative abundance of tissue autofluorophores like collagen, NADH and FAD [5, 10, 11]. These changes result from various cellular processes that are linked to oral dysplasia like increased NADH-FAD activity in the epithelial layer and degradation of stromal collagen. The ability of fluorescence imaging modalities to probe these biochemical changes makes them suitable for the diagnosis of oral cancer. Based on the source of contrast, fluorescence imaging techniques are often classified into two categories, namely, steady-state and time-resolved. There are two main advantages of time-resolved measurements over steady-state measurements. First, using fluorescence lifetime information from time-resolved measurements, it is possible to distinguish between multiple fluorophores that have overlapping emission spectra and are thus indistinguishable using only intensity information obtained from steady-state measurements. Second, unlike steady-state measurements, lifetime measurements

#261106 Received 15 Mar 2016; revised 18 Apr 2016; accepted 19 Apr 2016; published 25 Apr 2016

are more robust to variations in the intensity of excitation light and fluorophores' concentrations. Despite the aforementioned advantages of using lifetime information, most studies on the applications of fluorescence imaging for oral cancer diagnosis are based on steady-state measurements. This is partly because unlike steady-state fluorescence measurements, which require simple instrumentation and minimal computation, instrumentation and signal processing for lifetime based systems are more complex. Motivated by these observations, the first part of the present study seeks to: (a) assess whether using lifetime information, in addition to the fluorescence intensity information, offers any improvement in the diagnostic potential of fluorescence based imaging for oral cancer, and (b) examine the necessity of performing deconvolution (computational bottleneck in FLIM signal processing) for obtaining lifetime by evaluating the performance of an approximate method of lifetime estimation that does not require deconvolution.

Likewise, progression of oral cancer is also associated with several morphological changes in the oral epithelium, like epithelial thickening, loss of the layered structure, and irregular epithelial stratification [9, 12]. The high-resolution images of sub-surface tissue structures obtained from OCT imaging makes it a promising imaging modality for the diagnosis of oral cancer. In our recently published study [13], we described automated algorithms for OCT data processing for quantifying morphological features that are associated with the malignant transformation of the oral epithelium. In the second part of the current study, the diagnostic accuracy of these OCT features are compared with FLIM-derived biochemical features to investigate the advantage of using both biochemical and morphological information, as opposed to using only one type of information, for the diagnosis of oral cancer.

To evaluate the diagnostic potential of different features, a quantitative approach was taken in this study, in which several metrics (or *features*) describing various attributes of FLIM and OCT images were first obtained by using signal and image processing techniques. These features were subsequently used to train and test a statistical classification model. The relative performance of different feature sets in terms of their discriminatory ability was finally assessed by performing statistical tests on their classification accuracies.

2. Materials and methods

2.1. Imaging & data acquisition

In this section, we present a brief description of our previously developed combined FLIM-OCT system used to acquire the data used in this study. More details about the design and validation of the system can be found in [14]. The FLIM module of the multimodal system was implemented following a direct pulse-recording scheme, in which pixel rate could be equal to the laser repetition rate. A frequency tripled Q-switched Nd:YAG laser (SPOT-10-50-355, Elforlight Ltd.,England) was used as the excitation source (355 nm, 30 kHz maximum repetition rate, 1 ns pulse FWHM). A 50 μm multi-mode fiber was used to deliver the excitation light (shown in violet in Fig. 1) to the sample. The fluorescence emission from the sample (shown in blue) was directed through a 200 μm multi-mode fiber into a multi-spectral detection module, where it was separated into three spectral channels, namely, 390 ± 20 nm, 452 ± 22.5 nm, and 600 ± 125 nm, using a set of dichroic mirrors and filters. The emission from each channel was launched into fibers with different lengths (1 m, 10 m and 19 m) chosen to provide 45 ns interval between each emission band decay. The three emission bands, 390 ± 20 nm, 452 ± 22.5 nm, and 600 ± 125 nm, were selected to target endogenous fluorescence emission signal from collagen, NADH, and FAD, respectively. The three consecutive decays were detected with a MCP-PMT and sampled with a high bandwidth digitizer.

The Fourier-domain OCT module of the multimodal system was based around a 830 nm (40 nm FWHM) superluminescent light emitting diode (SLED) (EXS8410-2413, Exalos,

#261106 Received 15 Mar 2016; revised 18 Apr 2016; accepted 19 Apr 2016; published 25 Apr 2016

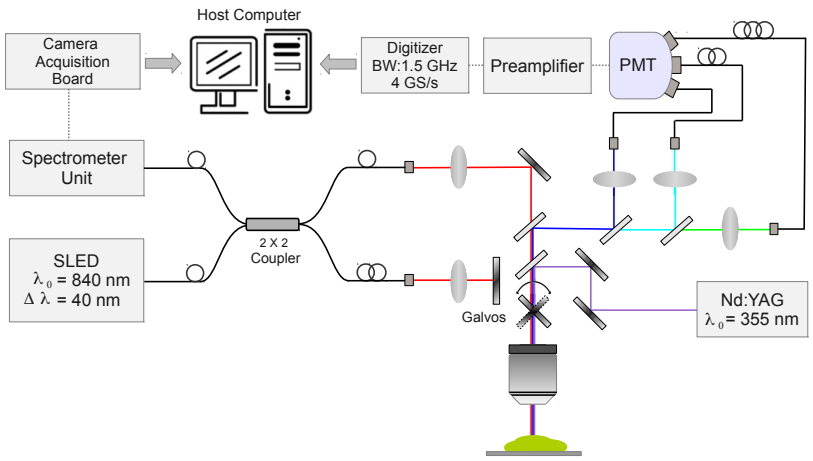

Fig. 1. Multimodal optical system for simultaneous co-registered OCT and FLIM imaging. The OCT beam is shown in red, the FLIM excitation beam in violet, and the FLIM emission beams in blue, cyan and green.

Langhome, PA) as the light source, providing an axial resolution of 7.6 μm (in air). Light from the SLED (shown in red) was directed to a 2×2 optical fiber coupler through a single-mode fiber, where it was split into reference and sample arms. The reflected beam from the reference mirror and the backscattered light from the sample were recombined at the fiber coupler and the spectral interferogram was obtained using a custom designed grating based high speed spectrometer (1200 lines/mm; bandwidth: 102 nm) and a CCD line-scan camera (line rate of up to 53 kHz). The detected signal was acquired and digitized using a high speed imaging acquisition board.

The OCT sample beam and FLIM excitation were combined via a dichroic mirror. Beam scanning over a field of view of 2×2 mm^2 (corresponding to $600 \times 600 \times 1024$ ($x \times y \times depth$) pixels for OCT and $60 \times 60 \times 768$ ($x \times y \times time$) pixels for FLIM) was achieved using a set of galvo mirrors. A broadly achromatic (UV-near IR) objective lens was used to simultaneously focus the OCT and FLIM excitation beams resulting in lateral resolutions of 13.4 μm for OCT and 100 μm for FLIM. The data acquisition module for the imaging system was developed in Labview and all processing was done in MATLAB.

2.2. *Animal protocol*

The standard Syrian golden hamster (*mesocricetus auratus*) cheek pouch model of epithelial cancer was used in this study. The animal protocol consisted of scheduled application of a suspension of 2.0% Benzo[a]pyrene (Sigma Aldrich Corporation, St Louis, Missouri) in mineral oil to the right cheek of 20 hamsters three times per week for up to 32 weeks. 11 control animals used in this study were similarly treated with mineral oil alone. The procedure was approved by the Institution for Animal Care and Use Committee (IACUC) at Texas A&M University. Before imaging, the hamsters were anaesthetized by an intraperitoneal injection of a mixture of ketamine and xylazine. The cheek pouches of the anaesthetized animals were inverted and positioned under the microscope objective of the imaging system. At the end of each scan, the imaged region was dried and the center of the area was marked with ink to allow correlation between the imaging and biopsy sites. After imaging, the animal was euthanized by barbiturate overdose. This ink mark was centered while extracting the imaged region for histopathological

#261106 Received 15 Mar 2016; revised 18 Apr 2016; accepted 19 Apr 2016; published 25 Apr 2016

processing using a biopsy punch tool with a diameter of 1 cm. In order to avoid smudging of the ink, sample was submerged in ethanol overnight before storing it in formalin.

2.3. Imaging & histological evaluation

Based solely on visual examination of cheek pouches, several imaging sites (potentially malignant sites in animals in the treatment group) were identified and imaged using our multimodality system. Biopsy samples from the imaged areas were processed following standard procedures for histopathology analysis (H&E staining). On an average, 10 sections per tissue sample were obtained and each section was assessed by a board certified pathologist to be one of the following five grades: (i) normal (G0), (ii) hyperplasia and hyperkeratosis (G1), (iii) hyperplasia with dysplasia (G2), (iv) carcinoma in situ (G3), and (v) squamous cell carcinoma (G4). Upon histopathological evaluation, the majority of the imaging sites were found to have mixed histology, such as cases showing both pre-cancerous and cancerous regions in the same sample. Since our classification method, like most standard machine learning algorithms, assume mutual exclusivity of classes, only a small subset of 48 samples out of a total of 153 samples for which the histopathological assessment was unambiguous (based on the criteria described next) could be used in our study. Specifically, for classification analysis, the following criteria was used to assign class labels to each tissue sample: (i) class 1 (benign; 22 samples): samples from the control group (15 samples) and samples for which all histology sections were graded as G1 (7 samples), (ii) class 2 (pre-cancerous; 12 samples): samples for which at least 50% sections were graded as G2 or G3 and none of the sections were graded as G4, and (iii) class 3 (cancerous; 14 samples): samples for which all sections were graded as G4.

2.4. Data processing

2.4.1. FLIM features

For a systematic analysis of FLIM features, we grouped the FLIM features into two groups, namely, the *exact* FLIM features and the *approximate* FLIM features, each containing nine features. Each group of nine features comprised the normalized fluorescence intensity and two lifetime features (which shall be referred to as the average lifetime and the $1/e$ lifetime) for each spectral channel (thereby making a total of 3 features $\times$ 3 spectral channels = 9 features). The average lifetime is defined as the average time a fluorophore stays in the excited state. Thus, if $h(t)$ denotes the time-resolved fluorescence decay, then the average lifetime, denoted by τ_{avg}, is estimated as: $\int_0^\infty t.h(t)\,dt / \int_0^\infty h(t)\,dt$. The $1/e$ lifetime provides another way of quantifying the rate of fluorescence decay, which is based on the assumption that the fluorophore follows first order decay dynamics. In such a case, the fluorescence lifetime can be estimated as the time at which the fluorescence intensity is reduced to $1/e$ of its maximum value. The $1/e$ lifetime shall be denoted by $\tau_{1/e}$. The average lifetime is different from $1/e$ lifetime in that it takes into account the entire fluorescence decay to estimate the lifetime, unlike $1/e$ lifetime, which is based on only the initial part of the fluorescence decay. Also, for a fluorophore that follows a first order decay kinetics, the two lifetimes, τ_{avg} and $\tau_{1/e}$, are equal.

In FLIM, the measured fluorescence decay (denoted by $y(t)$) is mathematically modeled as a convolution of the instrument response (denoted by $x(t)$) with the intrinsic fluorescence decay of the sample at that pixel (denoted by $h(t)$), *i.e.* $y(t) = x(t) * h(t)$. Thus, to obtain the intrinsic fluorescence decay for lifetime estimation, the instrument response needs to be deconvolved from the measured fluorescence decay. The average and $1/e$ lifetimes estimated from the deconvolved fluorescence decays, together with the normalized intensities constituted the exact FLIM features. The approximate FLIM features also comprised normalized intensities, average and $1/e$ lifetimes. However, unlike the exact lifetime features, which were obtained from the deconvolved decay profiles, the approximate lifetime features were estimated directly from

#261106 Received 15 Mar 2016; revised 18 Apr 2016; accepted 19 Apr 2016; published 25 Apr 2016

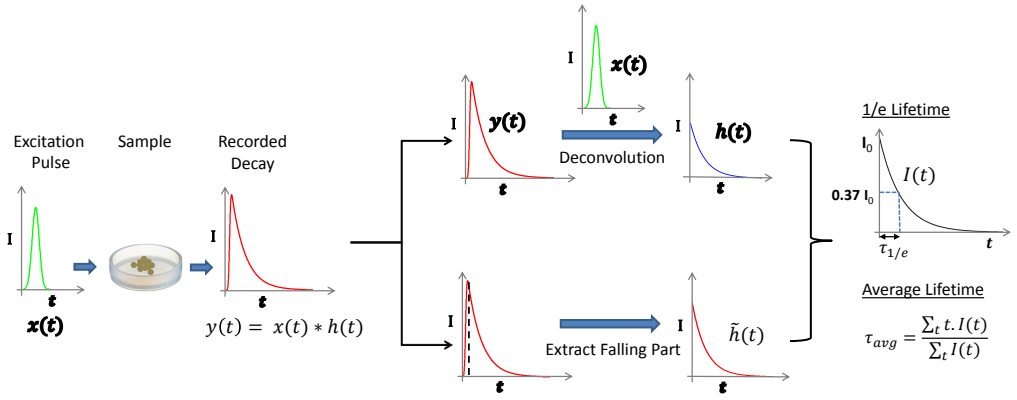

Fig. 2. Schematic illustrating the process of obtaining exact and approximate lifetime features. The exact lifetimes are obtained from the deconvolved intrinsic decay $h(t)$ shown in blue, whereas the approximate lifetimes ate obtained from the falling part of the recorded decay $y(t)$, denoted by $\tilde{h}(t)$, which is used as a substitute for the the intrinsic decay.

the undeconvolved recorded decay profiles by using the falling part of the recorded decay as a surrogate for the deconvolved decay profile, and subsequently estimating the lifetimes (both average and $1/e$) in a manner similar to the exact lifetimes. The process of obtaining the exact and approximate lifetimes is illustrated in Fig. 2.

As noted earlier, the 2-D FLIM feature maps contain 60×60 pixels over a FOV of 2×2 mm^2. For the analyses presented here, these feature maps were spatially averaged (window size: 3×3 pixels) to produce FLIM feature maps of size 20×20 pixels.

2.4.2. OCT features

We defined several features to characterize tissue morphology in OCT images. These features can be broadly categorized into two groups, which we shall refer to as A-line derived and B-scan derived features. Here, we briefly describe the the process of obtaining these features, which is also outlined in Fig. 3. For a more detailed description, interested readers are referred to our recent publication [13]. As can be seen in Fig. 3(a), an OCT B-scan of a normal oral tissue has a layered appearance, where different layers of the oral tissue can be seen as bright and dark bands. This is in contrast to the B-scan of a cancerous oral tissue, shown in Fig. 3(b), which has brightest intensity at the surface, which gradually fades off with depth. This layered versus non-layered structure is also evident by examining the filtered A-lines in a B-scan. In the case of layered tissue, an A-line has multiple prominent peaks corresponding to the different layers (Fig. 3(f)), whereas for an non-layered tissue, there is just one prominent peak at the tissue surface (Fig. 3(g)). To quantify these characteristics of A-lines, two types of A-line features were defined. The first one is referred to as the "peaks and valleys features" and the second one shall be called the "crossings features". To calculate the peaks and valleys features, the raw A-lined were first filtered using a nonlinear filtering process (details of which can be found in our recent publication [13]). Next, local maxima (peaks) and minima (valleys) of the filtered A-lines were detected and the following four peaks and valleys features were computed: (i) Σp_i, (ii) $\Sigma p_i - \Sigma v_i$, (iii) $\Sigma p_i + \Sigma v_i$, and (iv) $\Sigma(p_i - (v_i + v_{i+1})/2)$, where p_i and v_i denote the normalized intensity values of the ith peak and valley, respectively. More details about the filtering process used can be found in our To calculate the crossings features, for each filtered A-lines, a crossings vector of size 15×1 was defined. The $jth.$ element of the crossings vector denotes the number of times the A-line intersects an imaginary line drawn parallel to the x-axis

#261106 Received 15 Mar 2016; revised 18 Apr 2016; accepted 19 Apr 2016; published 25 Apr 2016

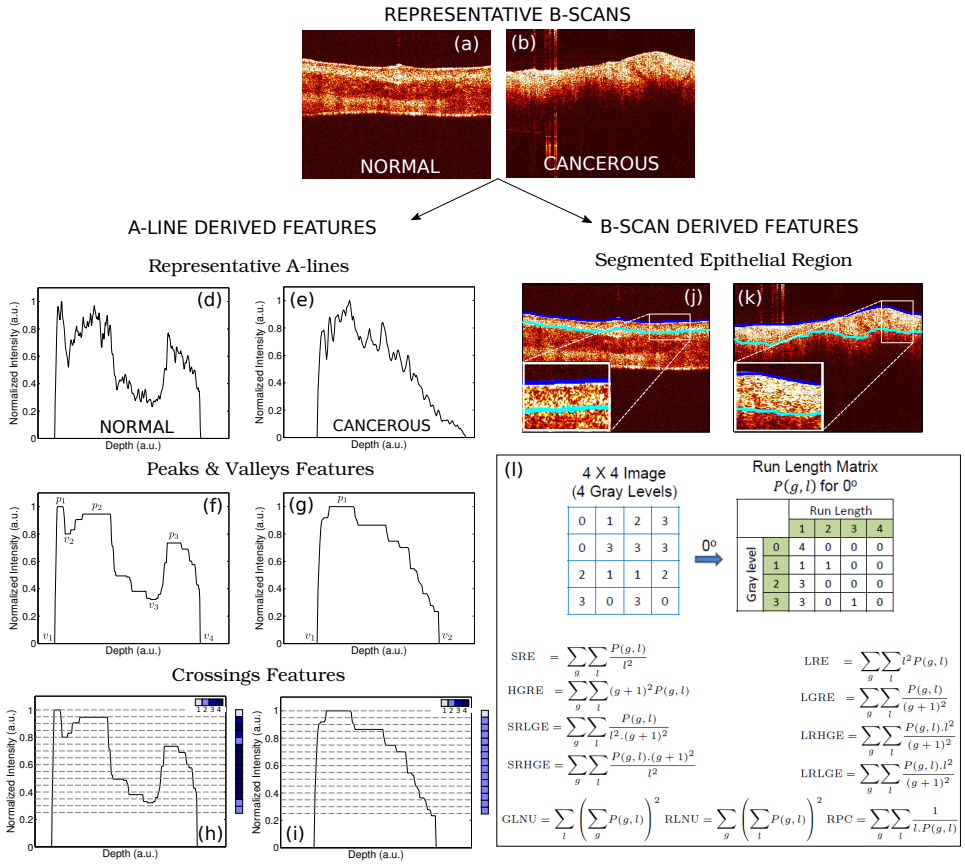

Fig. 3. Schematic illustrating the process of obtaining different OCT features. (SRE: Short Run Emphasis, LRE: Long Run Emphasis, HGRE: High Gray-level Run Emphasis, LGRE: Low Gray-level Run Emphasis, SRLGE: Short Run Low Gray-level Emphasis, LRHGE: Long Run High Gray-level Emphasis, SRHGE: Short Run High Gray-level Emphasis, LRLGE: Long Run Low Gray-level Emphasis, GLNU: Gray Level Non-uniformity, RLNU: Run Length Non-uniformity, RPC: Run Percentage)

(depth) and having a y coordinate (intensity value) of $(20-j)/20$. Intuitively, if an A-line has just one prominent peak, then all the elements of the crossings vector would be 2, as shown in Fig. 3(i), whereas, for an A-line that has multiple peaks and valleys, the elements of the crossings vector would be different than 2 (Fig. 3(h)). Four crossings features were estimated as the (a) mean, (b) median, (c) mode and, (d) standard deviation of the crossings vector. Overall, eight A-line derived features were obtained for each A-line, resulting in eight two-dimensional feature maps of size 600×600.

A different kind of structural information can be quantified by analyzing the texture of OCT images. Speckle in OCT imaging results from multiple scattering events and contains information about the size and distribution of the sub-resolution scatterers. The idea of using texture resulting from speckle is motivated by the fact that oral dysplasia is known to be characterized by basal cell hyperplasia and epithelial proliferation. The presence of dysplastic cells in the epithelial layer manifests as a highly speckled region in an OCT B-scan, which can be seen in the inset in Fig. 3(k). This is in contrast to the OCT image of a normal oral tissue shown in Fig. 3(j),

#261106 Received 15 Mar 2016; revised 18 Apr 2016; accepted 19 Apr 2016; published 25 Apr 2016

where different layers appear as more homogeneous bright and dark regions. Due to the multilayered nature of the oral tissue, the first step in computing these features was to segment out the epithelial region in a B-scan. In case of a layered tissue image, this was achieved by identifying the second prominent peak along A-lines in an OCT B-scan. In cases where layered tissue structure was absent, the boundary was detected based on segmenting the B-scan based on k-means clustering (details can be found in [13]). After delineating the epithelium boundary, the region between the surface and the epithelial boundary (region enclosed by the blue and cyan lines in Fig. 3(j) & (k)) was selected for texture analysis using gray level run lengths (GLRL) [15] to discriminate the uniform bright and dark banded region in an OCT scan of a normal tissue (inset in Fig. 3(j)) from an interspersed speckled region in the case of a cancerous oral tissue (inset in Fig. 3(k)). A "gray level run" is a sequence of consecutive pixels, along some direction, having the same gray level value. The length of a gray level run is the number of pixels in that run. To quantify texture based on GLRL, a two-dimensional GLRL matrix for a given direction was computed. The element (i, j) of the GLRL matrix denotes the number of times a run of length j and pixel value i is found in the image. The process of obtaining a run length matrix from an image is illustrated Fig. 3(l). To obtain texture features from a GLRL matrix, 11 different measures characterizing different properties of texture, like coarseness and non-uniformity, were computed. The GLRL features were computed for both vertical and horizontal directions denoted by 0° and 90°, and for two quantization levels, namely, binary and 32 gray levels, yielding 44 B-scan derived texture features. Both the A-line and B-scan derived OCT feature maps were spatially averaged (window size: 30×30) to yield a total of 52 two-dimensional OCT feature maps of size 20×20 (same as FLIM feature maps) corresponding to a pixel resolution of $100\,\mu$m/pixel.

2.4.3. Classifier design and evaluation

To evaluate the discriminatory ability of the OCT features, we used a popular classification algorithm called random forest. The choice of using random forests was based on its several attractive properties, such as: ability to handle more than two classes, low variance, robustness to noisy labels and, ability to work with large number of correlated features [16]. The performance of the classifier was assessed by using a variant of the leave-one-out cross-validation (LOO CV) method to estimate the classification accuracy. In the standard LOO CV procedure, all but one data-points are used for training the classifier and the left out data-point is used for testing. The process of training and testing is repeated in an iterative round-robin fashion, until all the data-points have been used as test data-points. The different iterative rounds of CV are called the CV *folds*. Since the data-points in our study correspond to pixels in 2-D feature maps, to avoid optimistically biased accuracy estimates resulting from spatial correlation between pixels, we performed leave-one-sample-out cross-validation (LOSO CV), in which the CV folds were performed over the datasets and not pixels. In addition to the mean classification accuracy, the sensitivity and specificity for each class was also computed.

2.4.4. Statistical comparison of different features sets

Two types of non-parametric statistical tests were used to compare the performance of different feature sets. To compare the performance of two sets of features, Wilcoxon signed rank test was used. The Wilcoxon test is a nonparametric procedure for comparing two sets of paired observations. The classification accuracies for different CV folds for classifiers based on the two feature sets were treated as the different observations. The null hypothesis for the test was that there there was no difference in the accuracies achieved by classifiers based on the two sets of features and the two-sided alternative hypothesis stated that it was otherwise. The second type of statistical test, namely, the Friedman test, was used to compare the performance of

#261106 Received 15 Mar 2016; revised 18 Apr 2016; accepted 19 Apr 2016; published 25 Apr 2016

Table 1. Results of Wilcoxon signed rank test for comparing the classification accuracies of the exact and approximate FLIM features ($n = 31, \alpha = 0.05$)

Feature Sets Compared	R^+	R^-	T statistic	z statistic	p value
Exact FLIM - Approx. FLIM	159	337	159	-1.7441	0.081

more than two feature sets. The Friedman test is the non-parametric equivalent of the ANOVA test. The null hypothesis for the Friedman test was that there was no difference in the accuracies achieved by classifiers based on the different sets of features being compared, and the alternative hypothesis stated that there was at least one pair of feature sets that had different classification accuracies. Briefly, the Friedman test ranks the different features sets being compared for each cross-validation fold based on the classification accuracies. The average ranks for the different feature sets are then used to compute a test statistic (χ_F^2 statistic) which is compared against a critical value determined based on a pre-determined significance level to check whether the measured average ranks are significantly different from the mean rank. If the critical value is less than or equal to the test statistic, the null hypothesis is rejected in favor of the alternative hypothesis. In such a case, a post-hoc test is conducted to identify which features sets are different. In this study, Holm procedure was used as a post-hoc test to compare the performance of the best feature set (as judged by the mean ranks obtained during Friedman test) against the remaining feature sets. Holm procedure compares a set of logically interrelated hypothesis of comparisons between the control and other feature sets. For each pairwise comparison, the p value is computed based on the mean rank difference between the two groups and the null hypothesis that there is no difference between the two groups is rejected if the p value is found to be less than an adjusted α value (called Holm α in this study).The choice of the Holm procedure was motivated by the simplicity and power of the test for performing multiple comparisons as described in [17].

3. Results and discussions

3.1. Comparison of the exact and approximate FLIM features

The mean classification accuracy for the exact and approximate FLIM features was found to be 83.2% and 80.9%, respectively. To test whether, in general, the exact FLIM features performed better than the approximate features, Wilcoxon signed rank test was performed. Out of the 48 pair-wise comparisons (corresponding to 48 LOSO cross-validation folds), the overall accuracy for the exact FLIM features was found to be the same as that for the approximate FLIM features in 17 comparisons, higher in 22 comparisons, and lower in the remaining 9 comparisons. The results of the Wilcoxon signed rank test ($n = 31, \alpha = 0.05$) summarized in Table 1 indicated that the classification accuracies for the exact FLIM features were not statistically different (p value = 0.081, $\alpha = 0.05$) than those for the approximate features. Although the mean accuracy for the exact FLIM features was higher than the approximate FLIM features, the p-value for the Wilcoxon test indicates that the difference between the overall accuracies for the two sets of FLIM features was not significantly different at $\alpha = 0.05$ level of significance. This suggests that an accuracy as high as that obtained by using the exact FLIM features can, in fact, be obtained by using the approximate FLIM features. The advantage of using approximate features over the exact features being that estimating the approximate FLIM features obviates the need for deconvolution and thus requires significantly less computational effort. In the following discussion, unless stated otherwise, "FLIM features" shall be assumed to refer to the "exact FLIM features".

#261106 Received 15 Mar 2016; revised 18 Apr 2016; accepted 19 Apr 2016; published 25 Apr 2016

Table 2. Results of Friedman test for comparing the classification accuracies of different FLIM feature sets ($n = 48, df = 3, \alpha = 0.05$)

	Feature Sets				χ_F^2 statistic	p value
	All	Int.	Int. & $\tau_{1/e}$	Int. & τ_{avg}		
Mean Rank	3.4	1.6	2.2	2.8	54.34	<0.001
Mean Accuracy	0.832	0.628	0.684	0.784		

Table 3. Post-hoc comparison table for Friedman test for comparing different exact FLIM features ($n = 48, \alpha = 0.05$)

Hypothesis Index	Feature Sets Compared	Mean Rank Difference	z statistic	p value	Holm α
1	All - Intensity	1.74	6.60	<0.001	0.017
2	All - Intensity & $\tau_{1/e}$	1.19	4.51	<0.001	0.025
3	All - Intensity & τ_{avg}	0.57	2.17	0.015	0.05

3.2. Comparison of intensity and lifetime features

To assess the improvement in the diagnostic accuracy achieved by using both intensity and lifetime information, as opposed to only intensity, the classification accuracies for the following four sets of exact FLIM features were compared: (i) All FLIM features (3 normalized intensities, 6 lifetimes), (ii) Intensities only (3 features), (iii) Intensities and $1/e$ lifetimes (6 features), and (iv) Intensities and average lifetimes (6 features). To test for statistical differences between the four sets of FLIM features, Friedman test ($n = 48, \alpha = 0.05$) was performed. The results of the Friedman test ($\chi_F^2 = 54.34, df = 3, p < 0.001$), summarized in Table 2, led to the rejection of the null hypothesis, suggesting that significant difference exists between the classification accuarcies of one or more feature sets. Based on the mean accuracies and ranks of the different feature sets summarized in Table 2, it was hypothesized that classification performance based on All FLIM features was better than the other feature sets. To test this hypothesis, Holm post-hoc procedure was conducted in which the classification accuracy of all FLIM features was compared with the other three feature sets. The results of the Holm procedure are presented in Table 3. Based on the results of the Holm procedure, all three hypotheses corresponding to the pairwise comparisons between All FLIM features and the other three feature sets were rejected suggesting that the performance of All FLIM features was statistically better that the other three feature sets.

These results are in agreement with the findings of an earlier study investigating the potential of time-resolved fluorescence measurements for oral cancer in a hamster cheek pouch model [18], where the authors found that using both intensity and lifetime features as opposed to only intensity features resulted in a significant improvement in the overall classification accuracy. The poor performance of intensity based imaging systems in oral cancer screening has also been reported in several other studies [19,20], where it has been noted that there is lack of substantive evidence to suggest that the use of autofluorescence intensity based imaging system improves the sensitivity and specificity of oral cancer beyond the conventional oral examination.

Also, the fact that using both lifetimes $\tau_{1/e}$ and τ_{avg} along with intensity results in a statistically higher overall accuracy compared to using either of them with intensity, suggests that

#261106 Received 15 Mar 2016; revised 18 Apr 2016; accepted 19 Apr 2016; published 25 Apr 2016

Table 4. Results of Friedman test for comparing the classification accuracies of FLIM and OCT features ($n = 48, \alpha = 0.05$)

	Feature Sets			χ_F^2 statistic	p value
	FLIM & OCT	OCT	FLIM		
Mean Rank	2.5	1.6	2.0	26.49	<0.001
Mean Accuracy	0.874	0.810	0.832		

Table 5. Post-hoc comparison table for Friedman test for FLIM and OCT features ($n = 48, \alpha = 0.05$)

Hypothesis Index	Feature Sets Compared	Mean Rank Difference	z statistic	p value	Holm α
1	FLIM & OCT - OCT	0.93	4.54	<0.001	0.025
2	FLIM & OCT - FLIM	0.51	2.50	0.006	0.05

using both lifetimes provide a more complete description of the fluorescence decay dynamics. A similar result was reported in [18], where the authors found highest overall classification accuracy for a combination of intensity, average lifetime and zeroth order Laguerre coefficient features, much like the combination of intensity, τ_{avg}, and $\tau_{1/e}$ in the present study.

3.3. Comparison of FLIM and OCT features

The mean accuracies and the ranks for the FLIM, OCT, and FLIM & OCT features are listed in Table 4. To test for statistical differences between the three sets of features, Friedman test ($n = 48, \alpha = 0.05$) was performed. The results of the Friedman test ($\chi_F^2 = 26.49, df = 3, p < 0.001$), summarized in Table 4, led to the rejection of the null hypothesis, suggesting that significant difference exists between the classification accuarcies of two or more feature sets. Based on the mean accuracies and ranks of the different feature sets, it was hypothesized that combining FLIM and OCT features leads to better classification accuracy than using either FLIM or OCT features. To test this hypothesis, Holm test was performed to compare the classification accuracy of the combined FLIM & OCT features with the individual OCT and FLIM features. Based on the results of the Holm procedure presented in Table 5, the two hypotheses corresponding to the pairwise comparisons between the combined FLIM & OCT features and the individual OCT and FLIM features were rejected suggesting that the classification accuracies for the combined FLIM & OCT features were statistically better that the individual OCT or FLIM features.

In addition to the mean classification accuracies, the sensitivity and specificity for each class (benign, pre-cancerous and cancerous) for the different feature sets were also computed. Based on the values listed in Table 6, it can be seen that the FLIM features had higher sensitivity and specificity than the OCT features for the benign class. On the other hand, OCT features had better sensitivity and specificity than the FLIM features for the cancerous class. For the pre-cancerous class, the two sets of features performed comparably. By combining both FLIM and OCT features, best sensitivity and specificity was achieved for all the classes.

The stacked bar graphs shown in Fig. 4, provide further insights into the classification performance. The graph shows the proportion of samples belonging to a given class: benign, pre-cancerous, or cancerous that were classified correctly or incorrectly. The bar graph in Fig. 4(a)

#261106 Received 15 Mar 2016; revised 18 Apr 2016; accepted 19 Apr 2016; published 25 Apr 2016

Table 6. Sensitivity and specificity for FLIM and OCT features

CLASS	Sensitivity			Specificity		
	Feature Sets			Feature Sets		
	FLIM & OCT	FLIM	OCT	FLIM & OCT	FLIM	OCT
Benign	0.882	0.868	0.786	0.920	0.934	0.884
Pre-cancerous	0.815	0.759	0.767	0.960	0.923	0.870
Cancerous	0.901	0.826	0.878	0.920	0.891	0.953

suggests that the classification based on the FLIM features provides relatively poor discrimination between the pre-cancerous and cancerous samples (cyan and red) compared to the classification based on the OCT features. A possible reason for this could be that the information contained in the fluorescence signal obtained from the samples in the pre-cancerous and cancerous classes was not discriminatory enough to achieve an accurate class separation, a fact noted by several other researchers investigating the potential of fluorescence based imaging methods for the diagnosis of oral cancer [18, 21, 22]. In contrast, significant difference in tissue morphology between the pre-cancerous and cancerous classes enabled superior performance of the OCT features.

Likewise, the bar graph in Fig. 4(b) suggests that the OCT features perform poorly compared to the FLIM features in discriminating benign samples from pre-cancerous samples (green and cyan). Following similar reasoning as before, this confusion between the two classes could have possibly resulted from the lack of sensitivity of the OCT features used in this study, to characterize differences in tissue morphology between samples in the benign and pre-cancerous classes. To overcome this limitation, more OCT features that are capable of characterizing morphological indicators of early dysplasia like changes in keratinization, epithelial thickening, irregular epithelial stratification and broadening of rete ridges could be evaluated and combined with the existing set of OCT features. Although a standard criteria of interpreting OCT B-scans for grading oral dysplasia has not been developed yet, previous studies on OCT-based diagnosis of oral cancer, both in hamster cheek pouches [6–8] and humans [9], have been successful in identifying the aforementioned histologically relevant morphological features in OCT B-scans, for diagnosing oral pre-malignancies and malignancies.

The classification based on both FLIM and OCT features takes advantage of the complimentary nature of the two feature sets to yield an improved performance compared to the classification based on either of those feature sets. This is indicated by the highest sensitivity and specificity values for the combined FLIM and OCT features, for all three classes.

It is of relevance to note that the performance of OCT-based classification is mainly limited by the nature of OCT features in terms of their ability to provide a sufficiently detailed description of tissue morphology. In this sense, one could expect to achieve a superior classification performance by using a set of more descriptive and richer OCT features, as discussed earlier. Moreover by using a longer wavelength (such as 1310 nm) OCT system, morphological information from deeper tissue structures could be obtained, which would possibly lead to further improvement in the diagnostic accuracy of OCT-based features. The general intensity and lifetime trends observed in our study were consistent with previously published studies, where fluorescence signal from non-cancerous tissue was strongest in the first emission band (390 ± 20 nm), while for cancerous tissue, the fluorescence intensity was highest in the second band (452 ± 22.5 nm) and lowest in the first band. This trend along with the lifetime values in

#261106 Received 15 Mar 2016; revised 18 Apr 2016; accepted 19 Apr 2016; published 25 Apr 2016

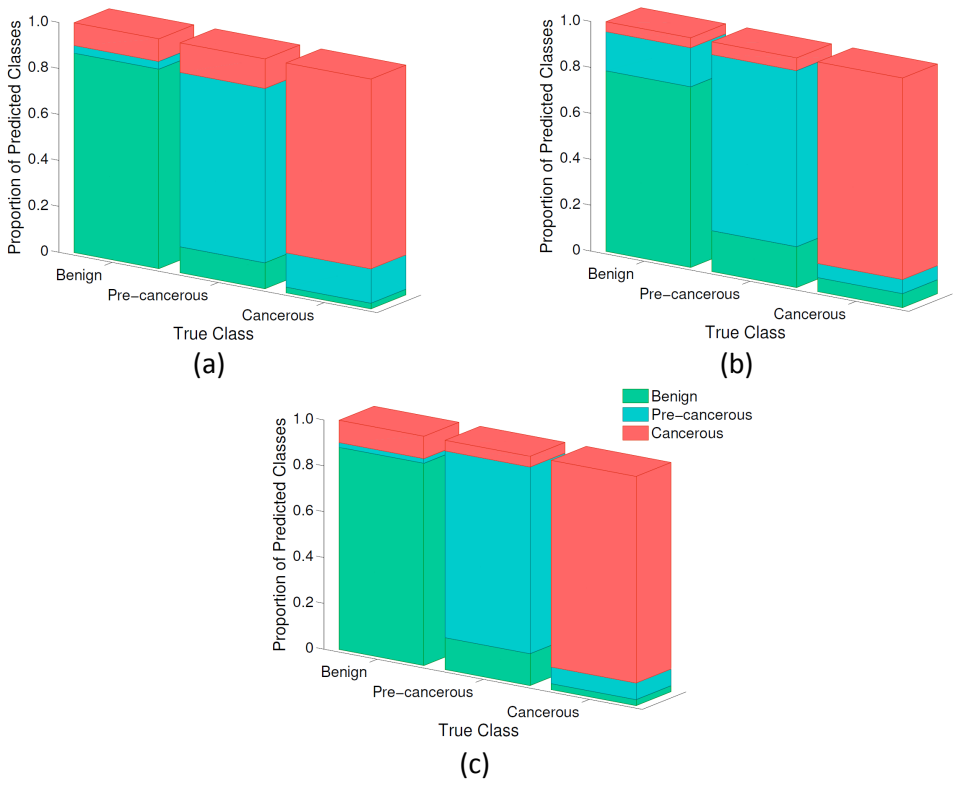

Fig. 4. Stacked bar graph showing the proportion of samples belonging to a given class: benign, pre-cancerous, or cancerous (x axis, true class) that were classified correctly (predicted as belonging to the same class) or incorrectly (predicted as belonging to one of the other two classes) for (a) FLIM features (b) OCT features (c) FLIM & OCT features.

the three bands suggests a relative decrease in collagen fluorescence and simultaneous increase in fluorescence from NADH and FAD with the progression of cancer. While this trend was true in general, particularly for the two extreme classes, namely, benign and cancerous, several exceptions were observed due to limitations of FLIM imaging arising from two main factors: (i) the difficulty to identify the specific sources contributing to the bulk fluorescence measurements, and (ii) the complex and sometimes similar nature of alterations in metabolic cofactors and structural proteins that occur in both benign and malignant oral conditions. For example, loss of stromal fluorescence is observed both in benign tissue with chronic inflammation caused by the breakdown of collagen crosslinks and in dysplastic tissue due to epithelial thickening and degradation of stromal collagen [6,23,24]. Similarly, an increase in keratin fluorescence can be observed in both benign and malignant oral conditions, due to the presence of keratotic lesions in both conditions [25,26]. This was also noticed in the current study, where histological evaluation of several samples in both the benign and pre-cancerous classes revealed the presence of hyperkeratosis. Although an increase in NADH fluorescence originating from the epithelial layer has been reported to be strongly correlated with neoplastic progression [4,24,27], in the presence of keratinized epithelium, most of the excitation light is back-scattered, which significantly reduces the fluorescence signal obtained from NADH. Moreover, the similarity in the spectral and temporal fluorescence characteristics of collagen and keratin [28,29] further

#261106 Received 15 Mar 2016; revised 18 Apr 2016; accepted 19 Apr 2016; published 25 Apr 2016

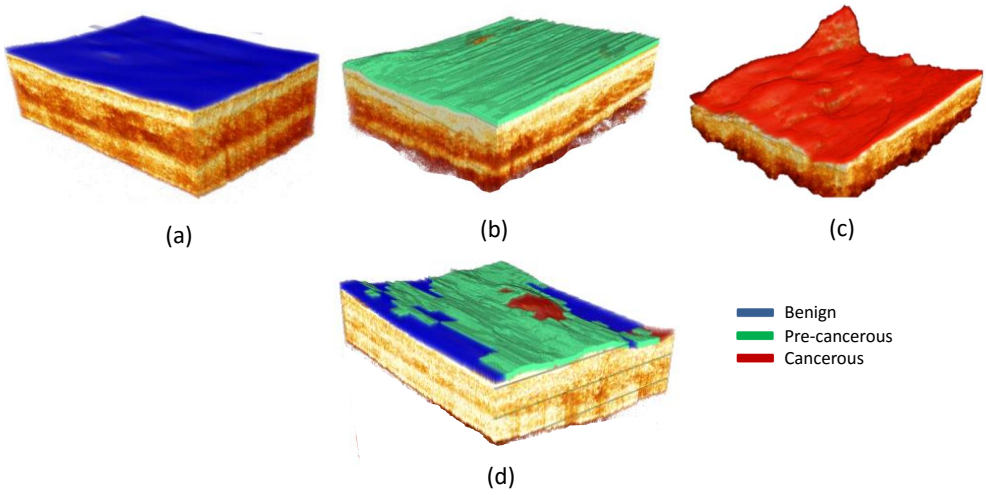

Fig. 5. Results of the classification presented as diagnostic maps overlaid on the OCT volumes for a (a) benign (b) pre-cancerous and (c) cancerous sample. Also shown is a representative sample that contains more than one class

confounds the FLIM-based diagnosis of oral dysplasia.

Another possible reason for the relatively inferior performance of the different feature sets for the pre-cancerous class could be the presence of mislabeled pixels (data-points) in the training data that belong to the pre-cancerous class. As mentioned earlier, a sample was labeled pre-cancerous if the histological evaluation of at least 50% of sections in that sample indicated some grade of dysplasia. This was done due to the lack of samples belonging to the pre-cancerous class that had sufficiently homogeneous evaluation across all the histology sections. As a result of this, not all the pixels in a pre-cancerous sample (although all labeled as pre-cancerous) were truly representative of pre-cancerous conditions. We speculate that the "label noise" arising in this way could also be responsible for the confusion between the pre-cancerous and other classes.

4. Conclusions

In this study, we demonstrated, perhaps for the first time, the feasibility of using automated quantitative image analysis algorithms for extracting relevant biochemical and morphological features from 3-D FLIM and OCT data to perform automated diagnosis of oral cancer. The results of the classification algorithm could be presented in form of a diagnostic map overlaid on the corresponding OCT volume for a better diagnostic interpretation as shown in Fig. 5. It is interesting to note in Fig. 5(d) that for regions classified as benign (color coded as blue) by the automated algorithm, the integrity of the layered tissue structures is intact, whereas for the regions classified as pre-cancerous (color coded as green), partial loss of the layered structure is evident.

The main finding to emerge from the present study was that the synergy of biochemical and morphological information obtained from FLIM and OCT, respectively, increases the sensitivity and specificity for detecting malignant and pre-malignant lesions, compared to using information from only one type of modality. The second major finding of our study was that

#261106 Received 15 Mar 2016; revised 18 Apr 2016; accepted 19 Apr 2016; published 25 Apr 2016

using both intensity and lifetime FLIM features, as opposed to only intensity features, provides significant improvement in the diagnostic accuracy. Moreover, to reduce the computational overhead of performing deconvolution in FLIM data processing, the performance of a new set of FLIM features, called the approximate FLIM features was also analyzed. The results of this analysis indicated that a diagnostic power as high as that of the exact FLIM features could in fact be obtained by using the approximate FLIM features. The advantage of using approximate features over the exact features being that estimating the approximate features does not require deconvolution and thus significantly reduces the computational effort.

We do recognize that due to the complex nature of the clinical and histological presentation of oral lesions, a further study using a larger pool of samples with more diverse histopathology is warranted to further validate the findings of the current study. Nevertheless, the findings of the present study provide an encouraging first step towards further exploration of automated analysis of multimodal imaging data for improved diagnosis of oral cancer.

Acknowledgments

This work was supported by grants from the National Institute of Health: R21-CA132433 and R01-HL11136.

#261106 Received 15 Mar 2016; revised 18 Apr 2016; accepted 19 Apr 2016; published 25 Apr 2016